論文 / 著書情報
Article / Book Information

題目(和文)	
Title(English)	Prosody Modeling Based on Gaussian Process Regression for Thai Speech Synthesis
著者(和文)	MoungsriDecha
Author(English)	Decha Moungsri
出典(和文)	学位:博士(工学), 学位授与機関:東京工業大学, 報告番号:甲第10923号, 授与年月日:2018年6月30日, 学位の種別:課程博士, 審査員:小林 隆夫,奥村 学,山口 雅浩,杉野 暢彦,篠﨑 隆宏
Citation(English)	Degree:Doctor (Engineering), Conferring organization: Tokyo Institute of Technology, Report number:甲第10923号, Conferred date:2018/6/30, Degree Type:Course doctor, Examiner:,,,,
学位種別(和文)	博士論文
Category(English)	Doctoral Thesis
種別(和文)	論文要旨
Type(English)	Summary

（博士課程）
Doctoral Program

論 文 要 旨

THESIS SUMMARY

専攻： Department of	物理情報システム 専攻	申請学位（専攻分野）： Academic Degree Requested	博士 (Engineering) Doctor of
学生氏名： Student's Name	Decha Moungsri	指導教員（主）： Academic Supervisor(main)	小林　隆夫
		指導教員（副）： Academic Supervisor(sub)	

要旨（英文 800 語程度）
Thesis Summary（approx.800 English Words ）

This thesis describes techniques to improve prosody modeling for Thai speech synthesis. Prosody modeling is an important issue, since Thai is a tonal language that possesses complicated intonation characteristics. Hidden Markov model (HMM)-based statistical parametric speech synthesis (SPSS) is one of popular speech synthesis frameworks and has been implemented in many languages, including Thai. The conventional HMM-based framework models acoustic features at the state level and uses tree-based context clustering to handle a variety of contextual factors. Although the conventional HMM-based SPSS framework can be utilized for prosody modeling, the synthetic speech is still imperfect. The degradation of synthetic speech quality is mainly caused by limitations of HMM-based SPSS and the suprasegmental level has not been appropriately incorporated into the prosody modeling. To overcome this problem, this thesis focuses on an alternative framework called Gaussian process regression (GPR)-based SPSS and proposes novel techniques to incorporate the suprasegmental level into the prosody modeling.

First, this thesis describes an implementation of GPR-based Thai speech synthesis including the GPR-based framework, Thai linguistics, and definitions of contextual factors, and kernel functions. The GPR-based SPSS was first introduced in Japanese speech synthesis to overcome the limitations of HMM-based SPSS. The GPR-based SPSS uses Gaussian process (GP) to model the relationships between frame-level contextual factors and acoustic features. GP is a nonparametric Bayesian model in which the model complexity grows as the amount of training data increases. The GPR-based SPSS uses a kernel trick that is more flexible than the tree-based context clustering in determining the similarity of complicated contextual factors. Another advantage is that the GPR-based SPSS framework models acoustic features at the frame level, which is more suitable than state-level modeling for acoustic features that change rapidly within one state. An experiment was conducted to evaluate the performance of the GPR-based framework by comparing it with the HMM- and DNN-based ones. The experimental result showed that the synthetic speech generated by the GPR-based method had more naturalness than the HMM- and DNN-based ones.

Secondly, this thesis describes novel techniques to incorporate suprasegmental features into the GPR-based duration and F0 modeling. The syllable level in the Thai language contains crucial linguistic functions such as stress, tone, and prominence that are primary factors of prosodic features. The conventional single-level model is insufficient for capturing these factors. To overcome the limitations of a single-level model, multi-level-model techniques were proposed for duration and F0 modeling. This thesis proposed two methods of multi-level duration modeling: two-stage prediction and the product of Gaussian process experts. Two-stage prediction uses phone and syllable-level duration models that are trained separately, and the predicted syllable duration from the syllable-duration model is used as an additional context in phone-level duration prediction. In the product of Gaussian process experts, the predictive distributions of phone- and syllable-duration models are combined by the product of experts framework. The mean of the combined predictive distribution is used as the predicted phone duration. This thesis examines multi-level F0 modeling by combining frame- and syllable-level models. F0 contours are generated by jointly maximizing predictive distributions of frame- and syllable-level models. The experimental results showed that the multi-model methods outperformed the single-model one.

Lastly, this thesis describes the use of stress information to improve prosody generation, because stress is a major factor that affects prosody. However, Thai is a non-lexical stress language whose stress cannot be obtained from the input text, and the manual labeling of stress information is time-consuming. To overcome such problems, an unsupervised technique was proposed to annotate the speech corpus with stress information. First, a dimensionality reduction technique, called the Gaussian process latent variable model (GP-LVM), was used to project acoustic features of stress onto a latent space in which the similarity between syllables can be easily observed. Then, an unsupervised clustering was performed to classify syllables into stressed and unstressed classes. The classification result showed that the use of the latent variables can achieve higher accuracy than the use of the observed acoustic features. In the experiment, the stressed/unstressed classes were used as an additional context in GPR-based prosody generation. Performance comparison of the GPR-based method was done using stress information obtained from manual labeling and unsupervised labeling. The objective evaluation showed that the prosodic features generated by the unsupervised labeling yielded a comparable result to the manual labeling technique. The subjective evaluation confirmed that the manual and unsupervised labeling methods produced similar quality of speech in terms of naturalness.

備考 ： 論文要旨は、和文 2000 字と英文 300 語を 1 部ずつ提出するか、もしくは英文 800 語を 1 部提出してください。
Note : Thesis Summary should be submitted in either a copy of 2000 Japanese Characters and 300 Words (English) or 1copy of 800 Words (English).